CLAIMS

1. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) generating a fuzzy set comprising, for word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

2. A method according to Claim 1, further comprising the step of:

(iii) adding a new document to said set of one or more documents and, using a set of words selected from said new document, performing an incremental update to said stored probabilities by means of steps (i) and (ii) performed in respect of said selected words using word sequences identified in said new document.

3. An information retrieval apparatus for use in retrieving information from a set of one or more documents, comprising:

an input for receiving a search query;

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from said set of one or more documents;

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities; and

information retrieval means for searching said set of one or more documents for relevant information using a received search query modified by said query enhancement means,

wherein said generating means are arranged, in use:

(i) for each word selected from said set of one or more documents:

(a) to identify, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

4. An information retrieval apparatus according to Claim 3, wherein said query enhancement means are arranged to identify, with reference to said generated set of probabilities, a word having a similar meaning to a term of said received search query and to modify said search query using said identified word.

5. An information retrieval apparatus according to Claim 3 or Claim 4, further comprising updating means for adding a new document to said set of one or more documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said generated set of probabilities in respect of words in said set of words.

6. An information retrieval apparatus for use in retrieving information in an information system, comprising:

an input for receiving a search query;

generating means for generating a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents;

query enhancement means for modifying a received search query with reference, in use, to said generated set of probabilities; and

information retrieval means for searching said information set for relevant information using a received search query modified by said query enhancement means,

wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

7. An information retrieval apparatus according to Claim 6, further comprising updating means for adding a new document to said sample set of one or more documents and, using a set of words selected from said new document and word sequences identified in said new document, performing an incremental update to said generated set of probabilities in respect of words in said set of words.

8. An information processing apparatus for use in an information processing apparatus, for use in an information system, for identifying information sets associated with a predetermined information category, the apparatus comprising:

generating means for generating, in the form of a matrix, a set of probabilities indicative of the semantic similarity of words selected from a sample set of one or more documents representative of the predetermined information category;

calculating means arranged to calculate, for each information set, a vector of values representing the relative frequency of occurrence, in the information set, of words represented in a matrix generated by the generating means; and

clustering means arranged to determine a measure of mutual similarity between pairs of information sets, using the respectively calculated vectors and the generated matrix, and to use the determined measures in a clustering algorithm to select one or more information sets to associate with the predetermined information category,

wherein said generating means are arranged, in use:

(i) for each word selected from said sample set:

(a) to identify, in documents of said sample set, word sequences comprising the word and a predetermined number of other words;

(b) to calculate a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) to generate a fuzzy set comprising, for groups of word sequences containing the word, corresponding fuzzy membership values calculated from the relative frequencies determined at step (b); and

(ii) to calculate, for each pair of words of said plurality of words, using respective fuzzy sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.

9. An information processing apparatus according to Claim 8, wherein the clustering algorithm is a hierarchic agglomerative clustering algorithm.

10. A method for determining the semantic similarity of words in a plurality of words selected from a set of one or more documents, for use in the retrieval of information in an information system, comprising the steps of:

(i) for each word of said plurality of words:

(a) identifying, in documents of said set of one or more documents, word sequences comprising the word and a predetermined number of other words;

(b) calculating a relative frequency of occurrence for each distinct word sequence among word sequences containing the word; and

(c) generating, from the relative frequencies determined at step (b), a set of probabilities representative of the contexts in which the word occurs; and

(ii) calculating and storing, for each pair of words of said plurality of words, using respective probability sets generated at step (i), a probability that the first word of the pair is semantically suitable as a replacement for the second word of the pair.